Game Data Analysis Report

Michael Dowd

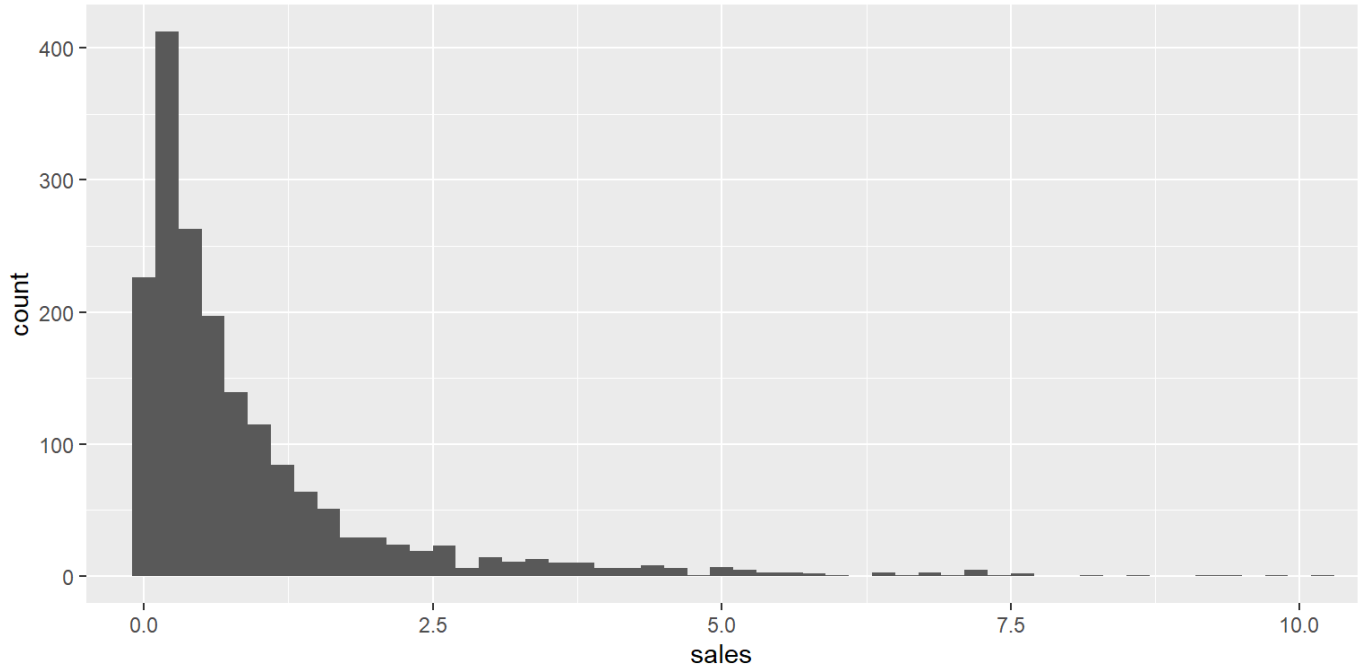
Problem Definition

The task is loosely defined - you're given a small dataset which is a subsample of game sales figures and you need to analyse the data and then, based firmly in the analysis, make recommendations to a game production company with regard to marketing and business strategy.

Dataset Exploration

Game Sales

The target variable I want to maximise is game sales. This density plot shows the overall range of sales values that have been observed in the data.



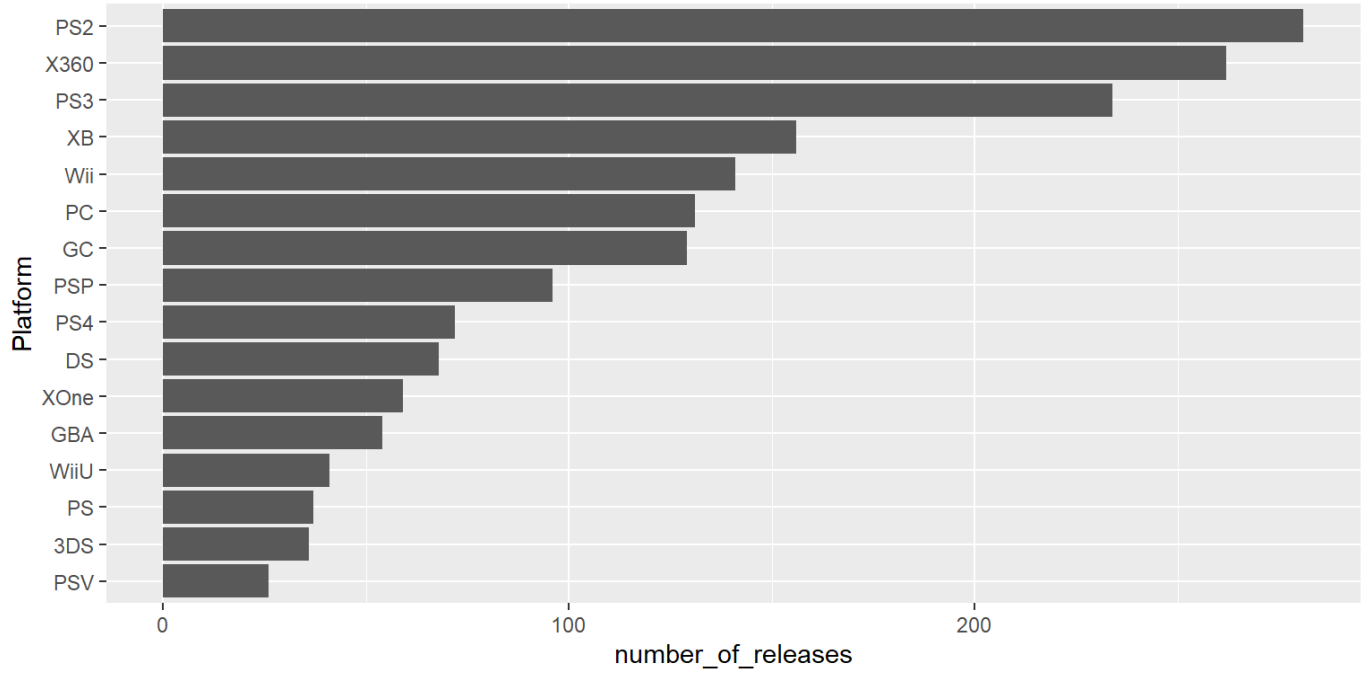
The data is right-skewed. There are a few extreme outliers and also there is a hard lower limit of 0 for sales while there is no upper limit. Here are some summary statistics:

Mean Sales	Standard Deviation	Median Sales	95th Quantile	Max Sales
1.134	2.993	0.52	3.76	82.53

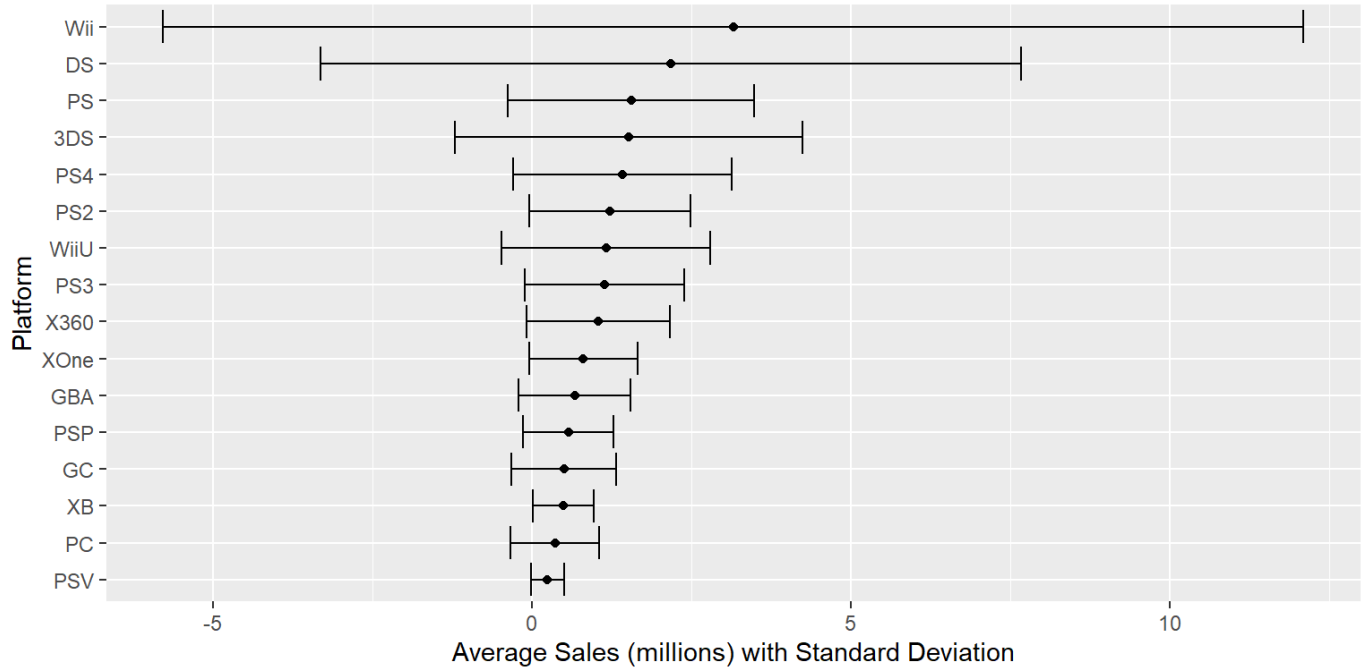
The median (520,000) is significantly smaller than the mean (1.13 million). Another interesting effect of the right-skew is that though the maximum sales for one game was 82.5 million, 95% of games sold 3.76 million or less.

Platform

The most popular platforms in the dataset with regards to number of games released are the PS2 and X360. These are both obsolete now and in fact the first current platform, the PS4 is number 9 on the list.

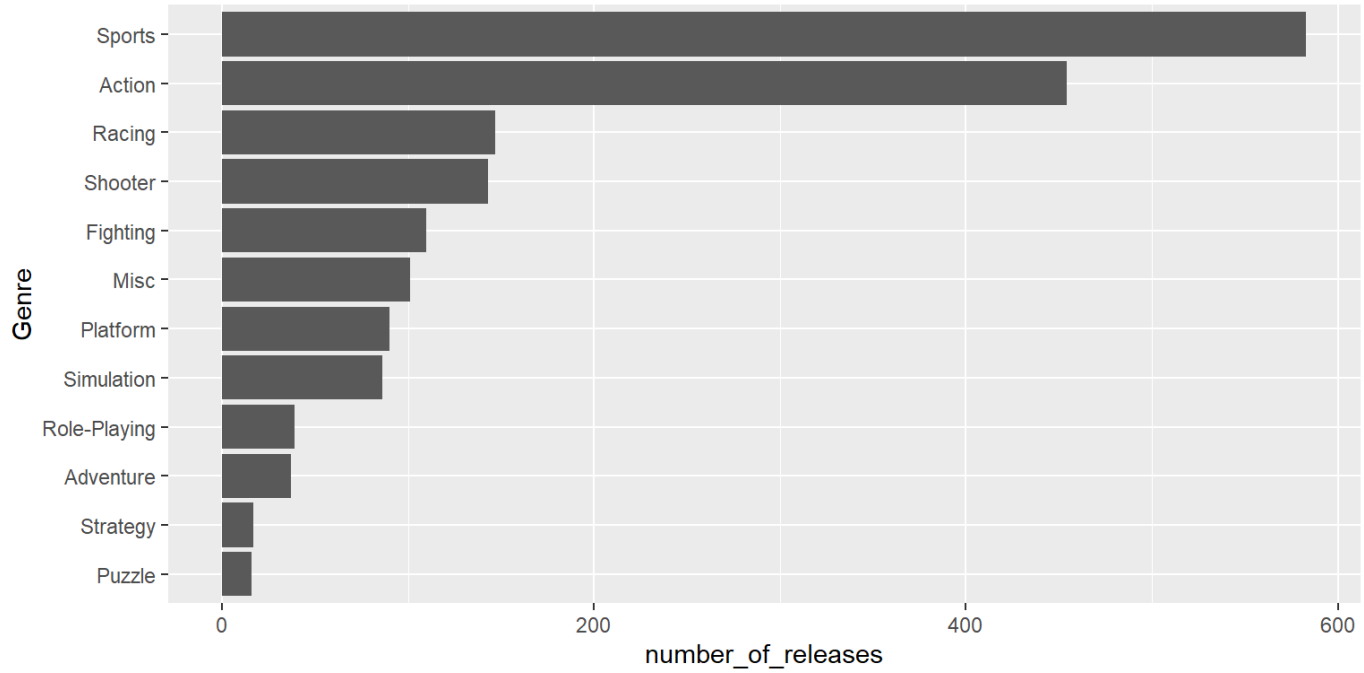


The average sales for each platform are shown below The standard deviation gives a visual guide to how much the sales for each platform vary.

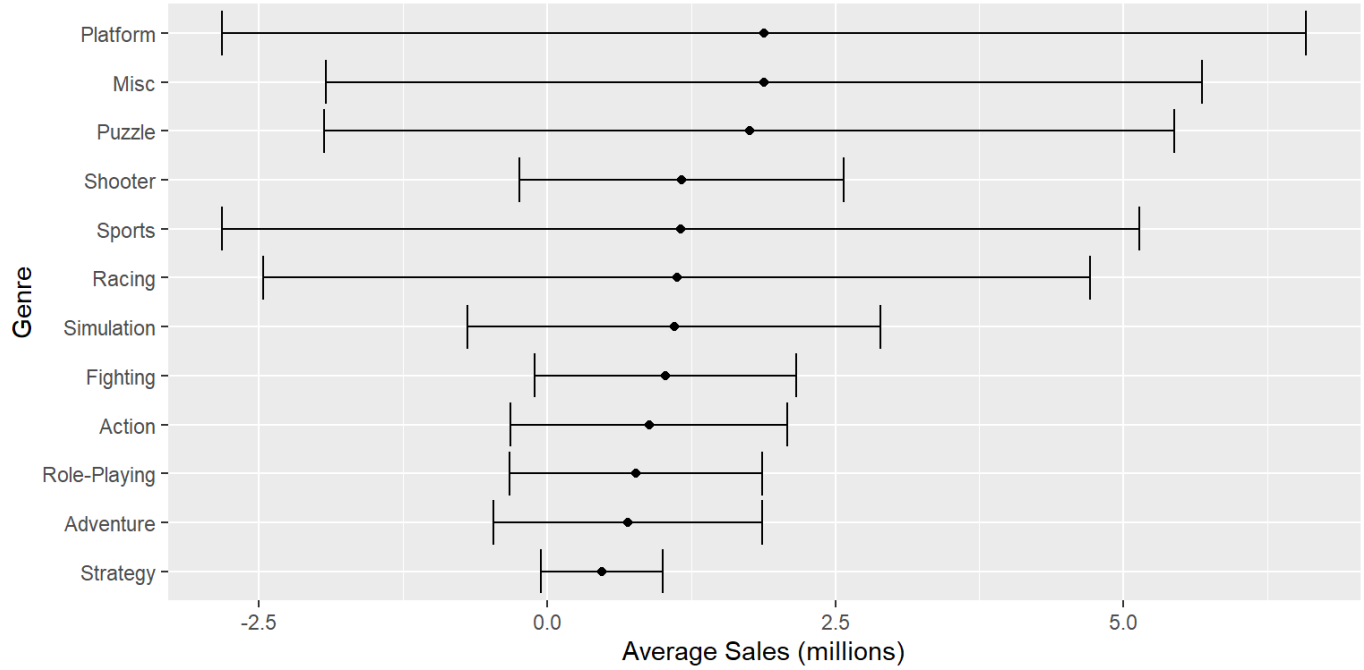


Genres

The most popular genre with regard to number of games release is sports, while puzzle has the least amount of games.



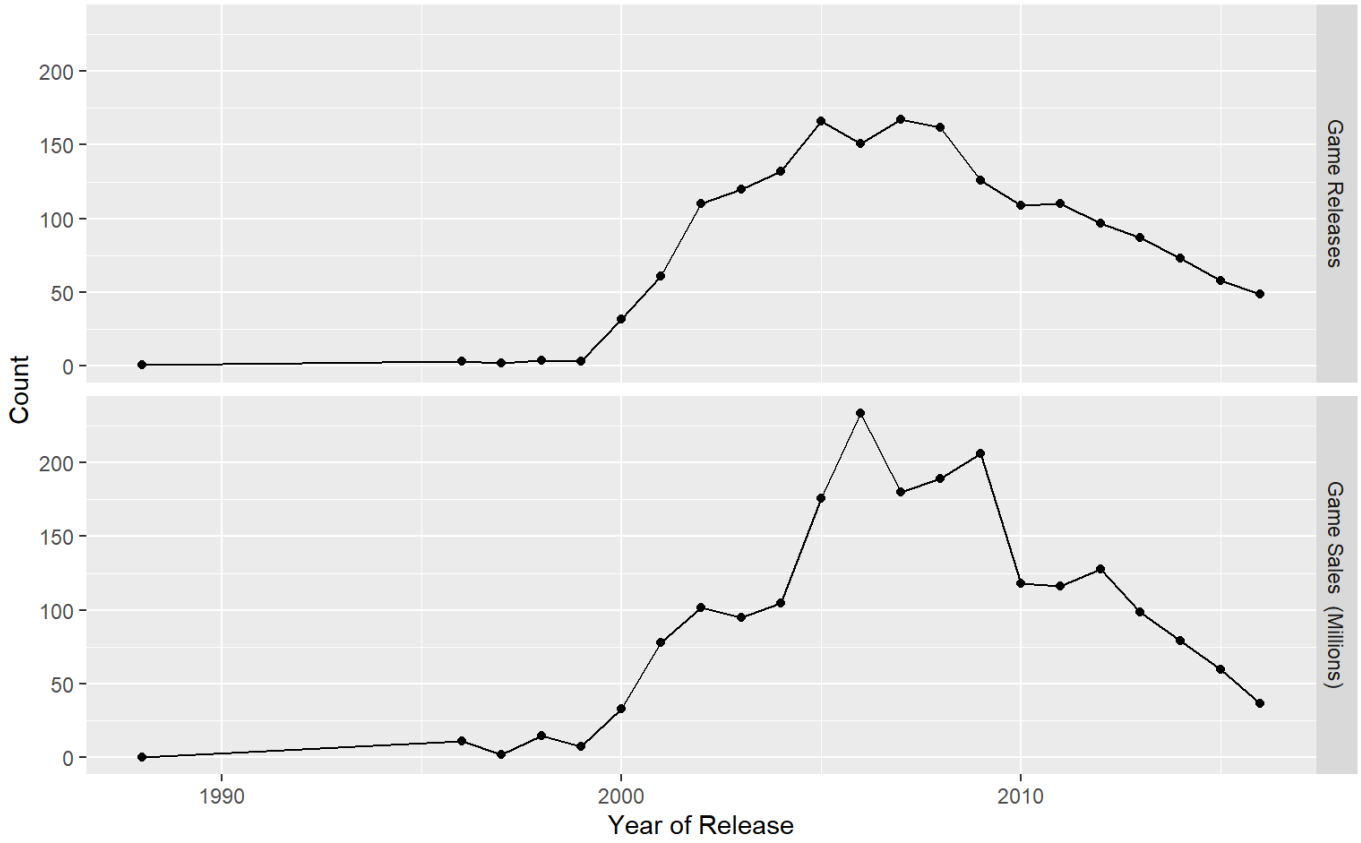
The average sales and standard deviation for each genre is shown below



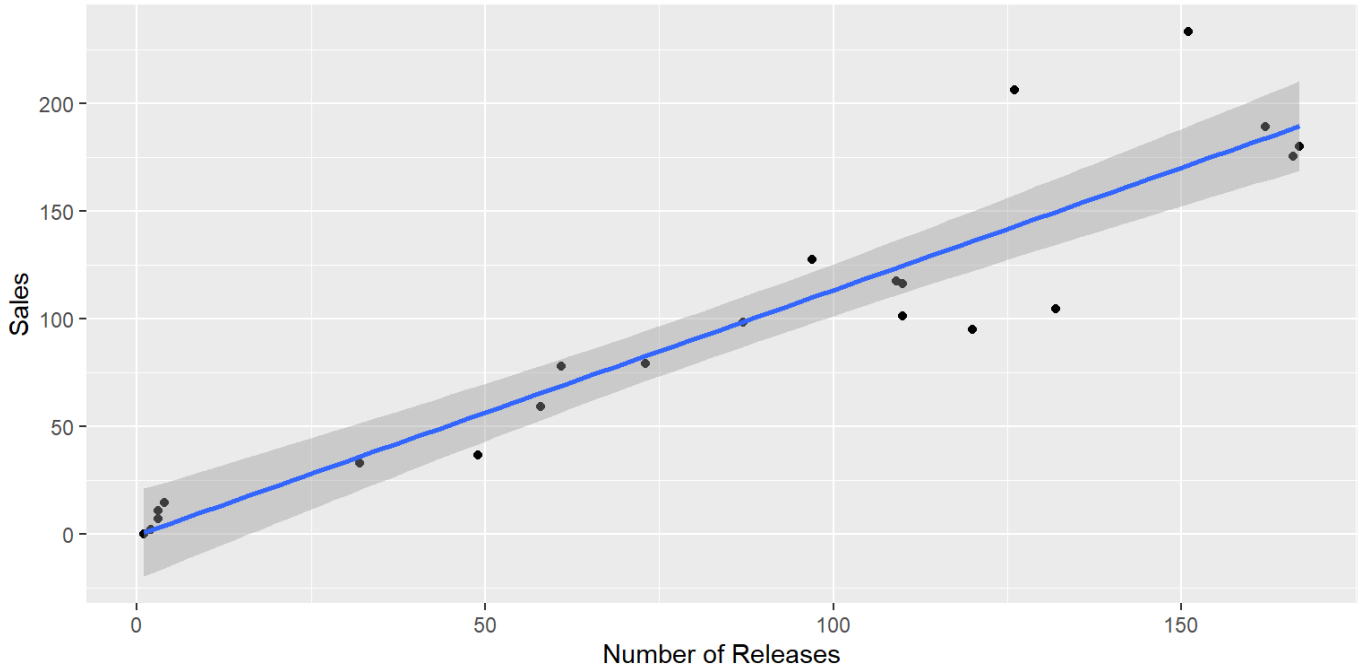
Interestingly, this chart shows the puzzle genre as having a high average sales figure, but large standard deviation, indicating that there is a high variance in this genre

Game Releases and Sales over Time

The chart below tracks the number of games released each year alongside the number of sales each year.



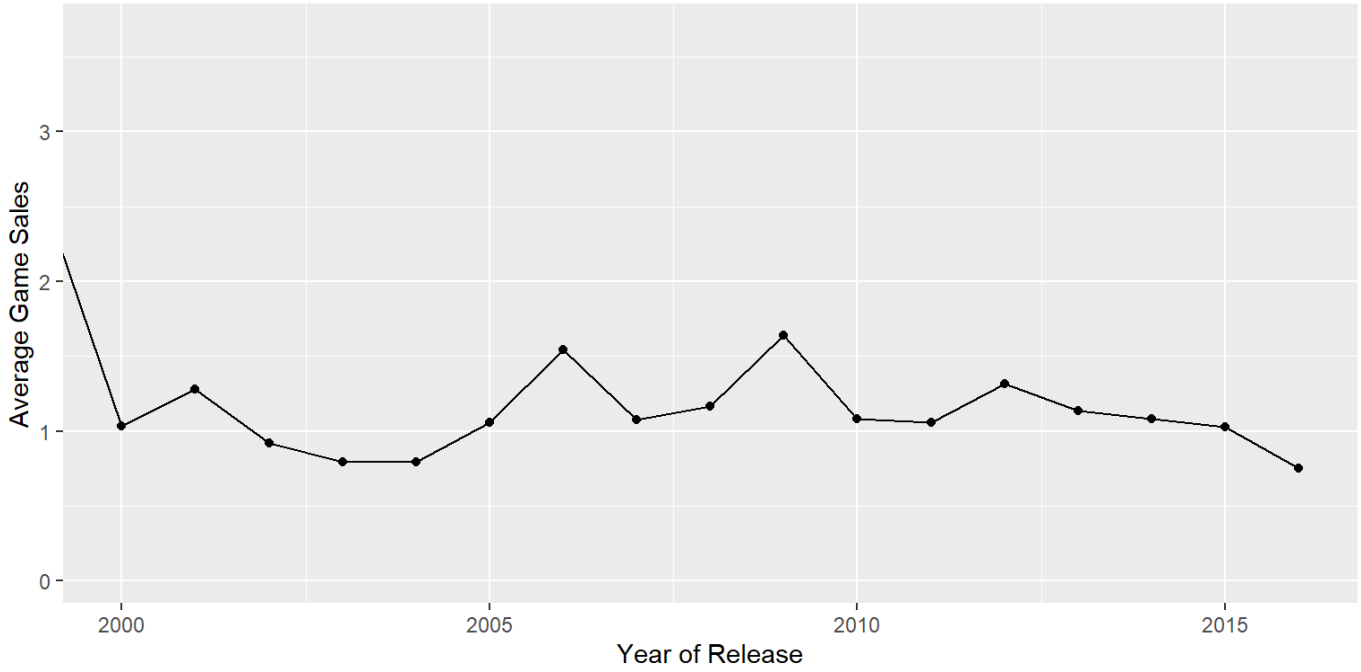
Interestingly the two series appear to track each other closely. To confirm this the correlation of the two series is checked:



Pearson's Correlation	p-value
0.933	2.483e-10

The correlation value is high and there is a small p-value, suggesting that if the null hypothesis were true (ie. no correlation exists) then we would almost never see these results. Therefore we can safely reject the null hypothesis and conclude there is a high correlation between these two series.

Though game sales through the years have changed drastically, this is accompanied by correlated changes in total sales, suggesting that perhaps average game sales each year aren't changing significantly. This is investigated below:



It is an important for a games company to know whether the average game sales are trending in any particular direction as this will give an indication of the market for future games.

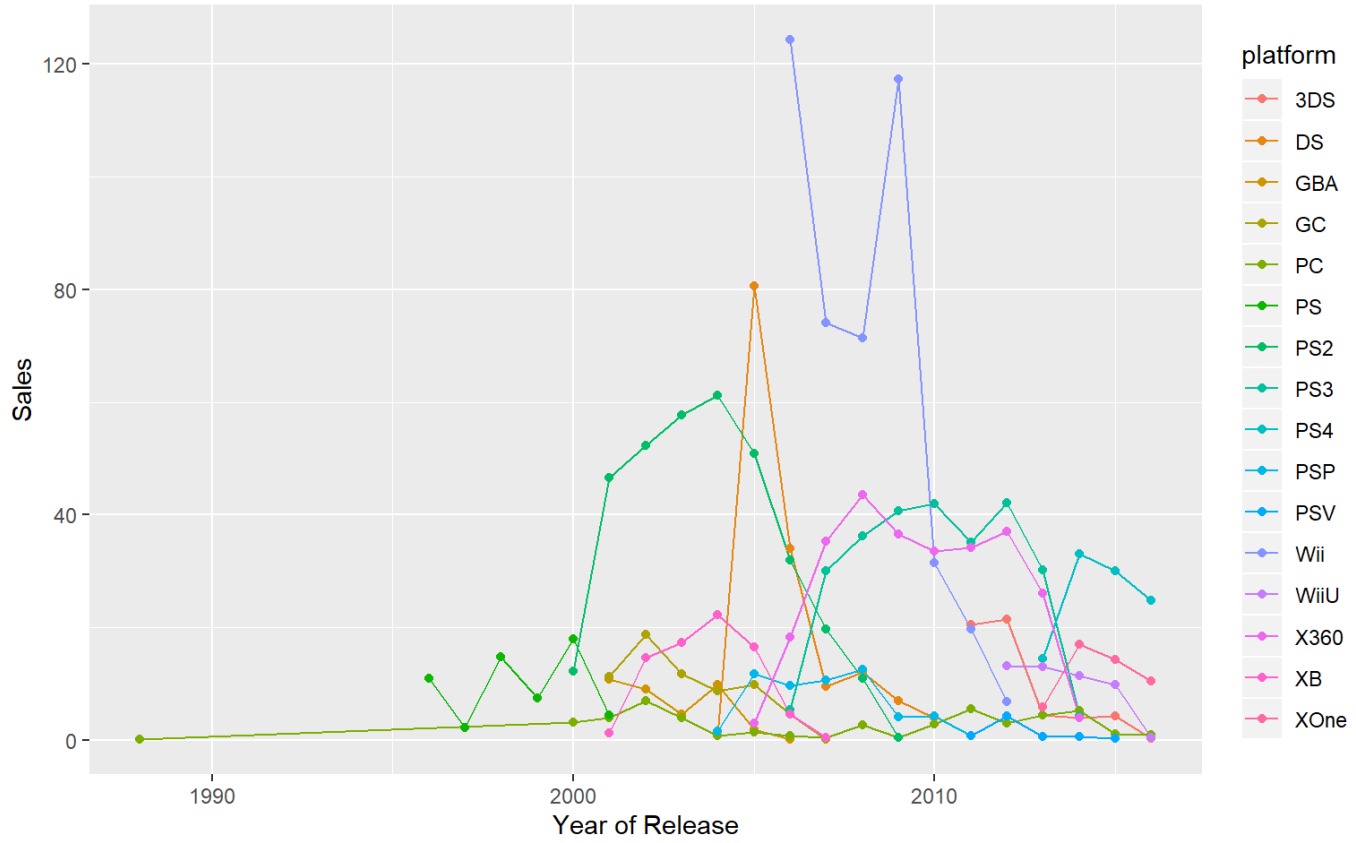
To formally test this I have used an anova test with year as the categorical variable and mean sales as the target variable. I also filtered out years prior to 2000 as there were very few sample points having large effect on the mean.

Variable	**p-value**
Year of Release	0.184

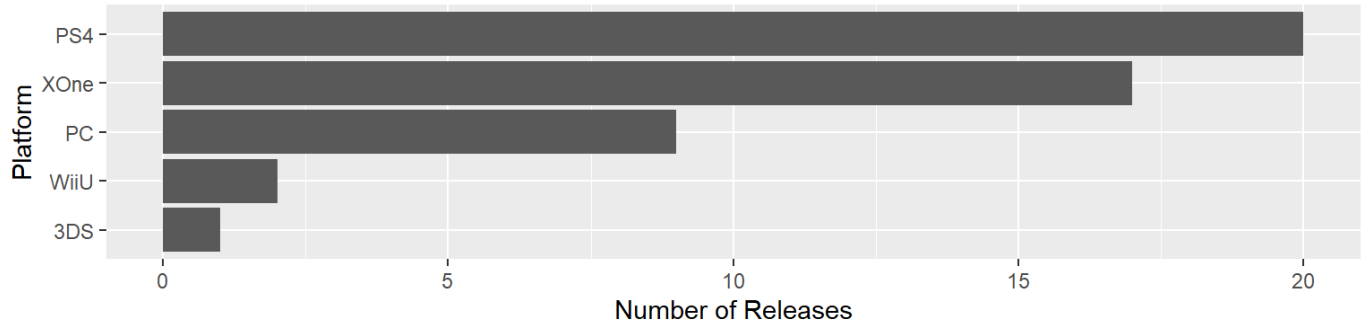
As we have a P-value > 0.05, there is not enough evidence to suggest that year of release is having any statistically significant impact on average game sales.

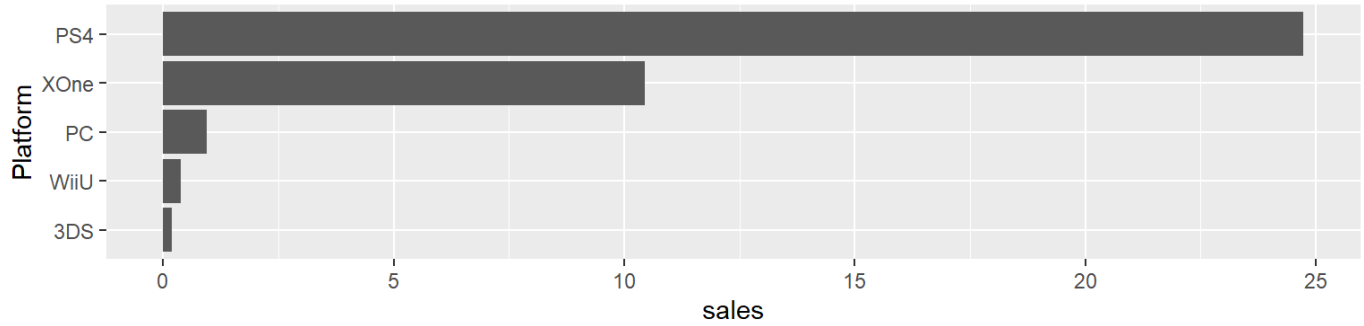
Platforms over time

An important property of the gaming industry is that gaming platforms change and become obsolete over time. The chart below shows the number of sales on each platform over time.



In the year 2016, games only sold on the following 5 platforms: 3DS, WiiU, PC, XOne and the PS4. The following charts show the breakdown for each platform in terms of number of games released and total sales.





There are a limited number of platforms to release on in 2016 and this will be a constraining factor in later suggestions.

Correlations and linear models

The aim of the task is to find variables correlated with higher sales. I will use a multi-variate linear model and Pearson's correlation to identify relationships between variables and the sales figure.

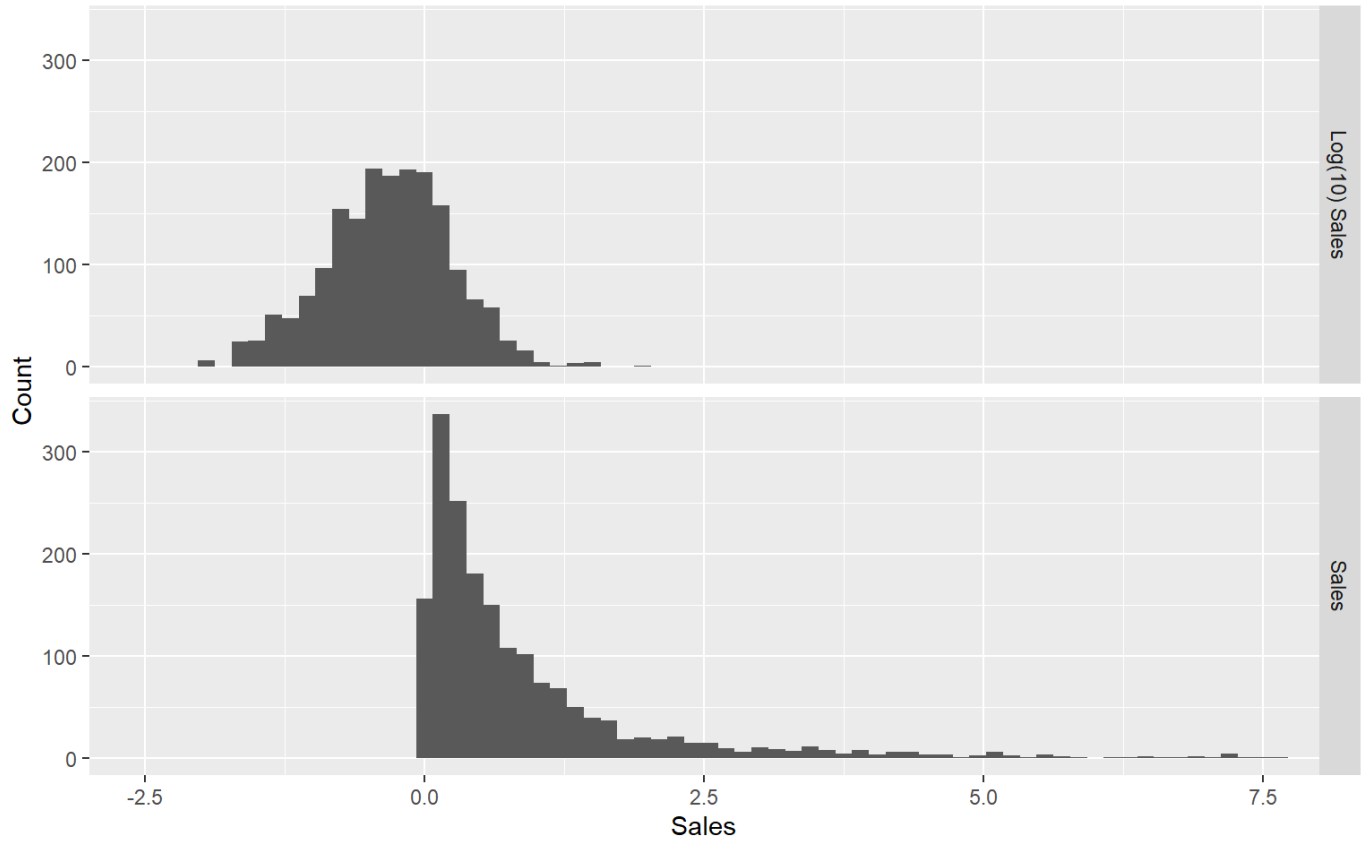
These two methods are designed to identify ‘linear’ dependencies between variables, eg. for variable x, there exists fixed values of A and B such that Sales = A+Bx. In fact this is a very strong assumption to make in this case, especially because:

- The sales data is not normally distributed (exhibits heavy right-skew)
- There is a hard cut-off at 0 (with many samples bunched around there)

In the case of game sales, where we are seeing the sales for some games being orders of magnitude larger than others, we are likely looking for is a non-linear exponential relationship, in other words, an increase in a variable x causes an exponential increase in game sales. To model this relationship I will investigate the log-adjusted sales and the actual sales as targets for linear relationships

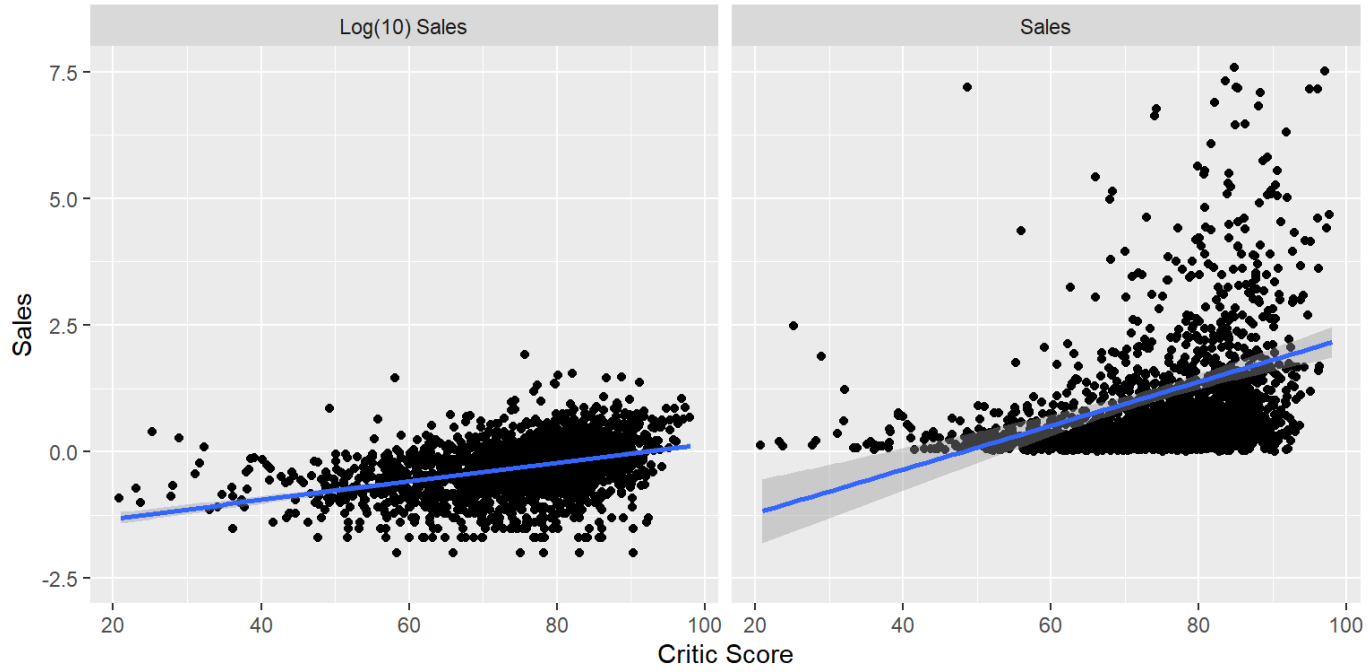
The below histograms show the distribution of non-adjusted and log-adjusted sales figures. The log-adjusted sales are in fact normally distributed and much more likely to exhibit linear relationships with other variables.

```
## Warning: NaNs produced
```



Game Critic Scores Analysis

I will now examine the critic review scores against the log-adjusted and non-adjusted sales figures.



The linear regression line here for the non-adjusted sales on the right is clearly less accurately fit, especially for games that received a lower critic score. The bands on either side of the fit line give an indication of the quality of fit and are much narrower for the log-adjusted sales data on the left.

To confirm this I've calculated the p-value for intercept and slope for both adjusted and non-adjusted sales below:

Non-adjusted sales critic score linear model

Variable	Value	p-value
Intercept	-2.084	2.37e-06
Critic Score Slope	0.043	2.04e-13
R-Squared	0.029	

Log-adjusted sales critic score linear model

Variable	Value	p-value
Intercept	-1.692	2e-16
Critic Score Slope	0.01845	2e-16
R-Squared	0.147	

The p-value for the slope in both cases is extremely low, so there is almost certainly a relationship here and these models are both statistically valid. However the P-values for the log-adjusted sales model is lower, as well as the R-Squared value being much higher, so this would lead me to conclude that the log-adjusted model is a better fit for this data.

The Pearson's correlation coefficient can also be tested against log adjusted and non-adjusted sales figures

Non-adjusted sales and critic score correlation	p-value
0.1709	2.039e-13

Log-adjusted sales and critic score correlation	p-value
0.3834	2.2e-16

In both correlation tests there is a very small p-value suggesting a high certainty of the results. The correlation between non-adjusted sales and critic score is relatively low, yet not very small, at 0.1709, while the correlation between log-adjusted sales is more than twice as strong at 0.3834.

Log-adjusted sales and user score correlation	p-value
0.2069	2.2e-16

Non-adjusted sales and user score correlation	p-value
0.09984	1.947e-05

Interestingly there are much smaller correlations between sales data and user scores, suggesting that purchasers of games tend to rely more on critic scores than user scores when making their decisions.

Multivariate Model

Platform, Genre, Rating & Publisher

I'm going to investigate these four variables together for the important reason of these being the variables that the game designer has direct control over.

Assumptions:

1. This analysis is for is a game developer, therefore it is meaningless to develop a model which relates sales back to the developer variable, as this variable can't be changed
2. The company can pitch a game 'profile' to a preferred publisher, so Publisher is a variable over which we have some control and it will be useful to know what publishers are related to high game sales.
3. The possible platforms are limited to PS4, XOne, PC, WiiU and 3DS. We might also consider developing for the soon to be released Nintendo Switch if there is good evidence to support this decision

The following code section executes backwards selection with p-values and comments indicated the eliminated variables and why they were eliminated

Non-adjusted Sales Linear Model	Genre	Publisher	Rating	Platform
P-Values	0.015545	2.2e-16	0.001175	2.2e-16

Log-adjusted Sales Linear Model	Genre	Publisher	Rating	Platform
P-Values	1.392e-09	2.2e-16	6.244e-12	2.2e-16

Two models are tested, one with log-adjusted sales and the other without adjusting the sales figures. In both models the P-Values for each of the 4 selected variables are less than 5% indicating that each variable statistically significant effect on the sales and none had to be eliminated.

R-Squared Comparison	Log-Adjusted Sales	Non-Adjusted Sales
R-Squared Value	0.3692	0.2433

The p-values in the log-adjusted model are much lower, especially for the genre and platform variables. Also the R-Squared value for the log-adjusted sales model is quite a bit higher than the non-adjusted sales model. The log-adjusted sales model is therefore a better fit for the data and I will proceed using log-adjusted sales.

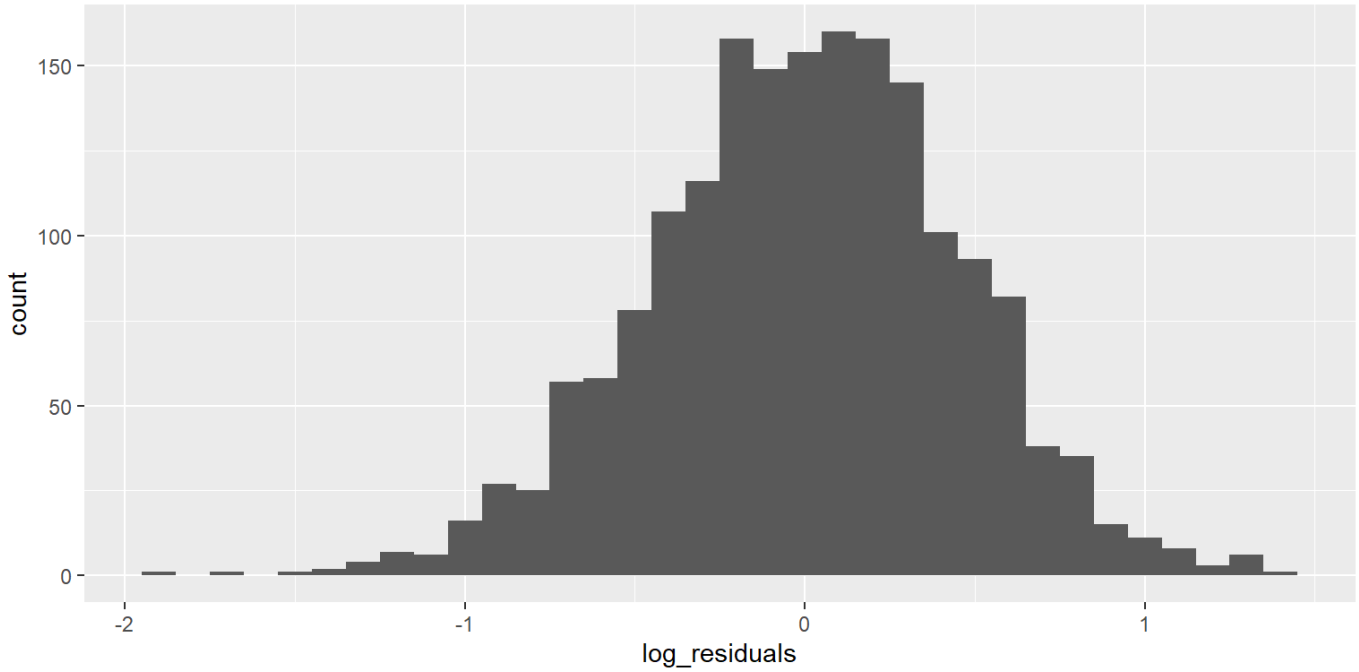
Linear Model - Checking Assumptions

1. Numeric variables are linearly related to target.

As there are no numeric variables being used in this model, nothing needs to be done here

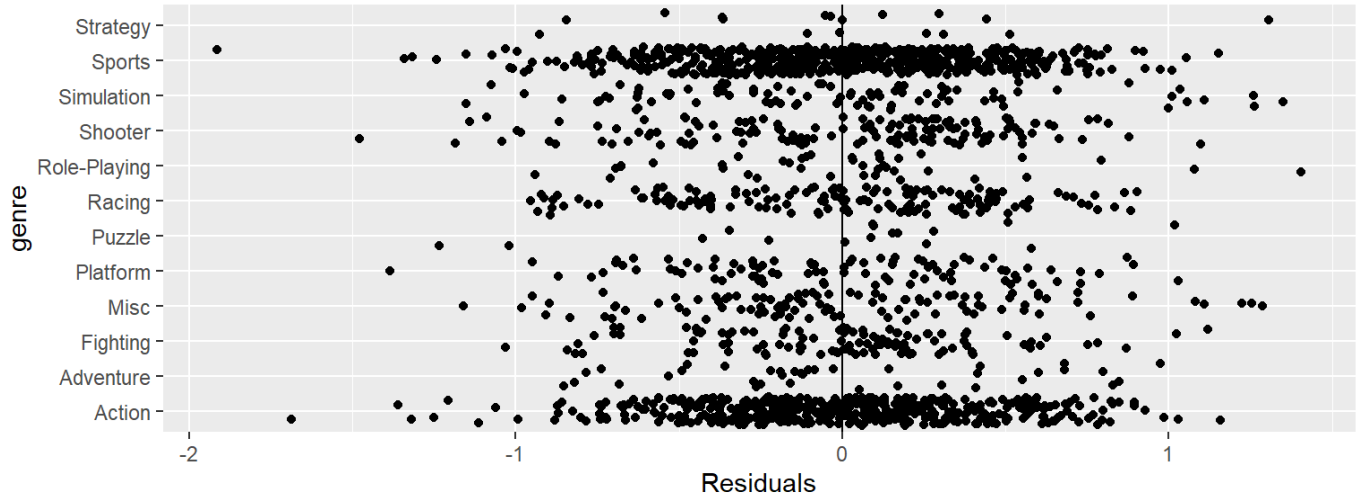
2. Normality of Residuals

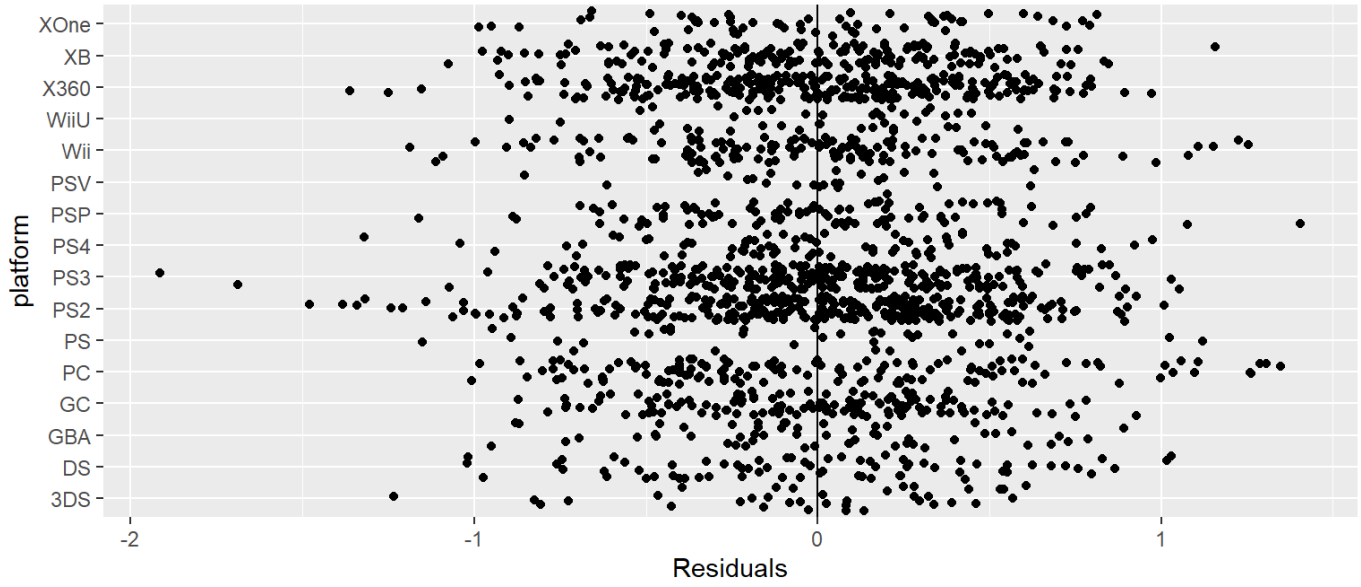
The residuals in the log-adjusted sales model are approximately normal and satisfy this assumption.

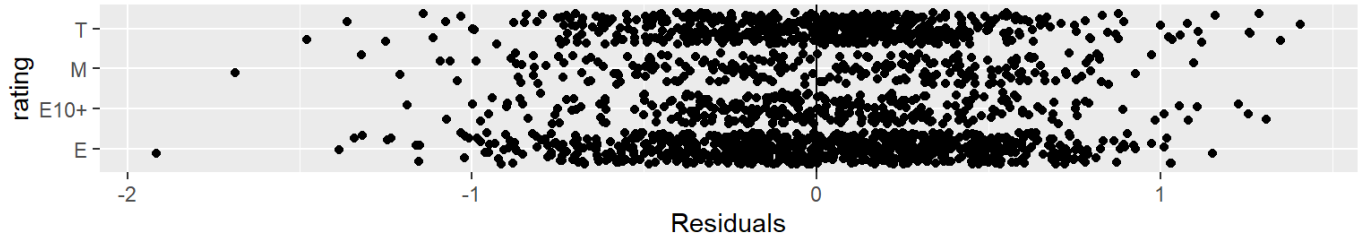


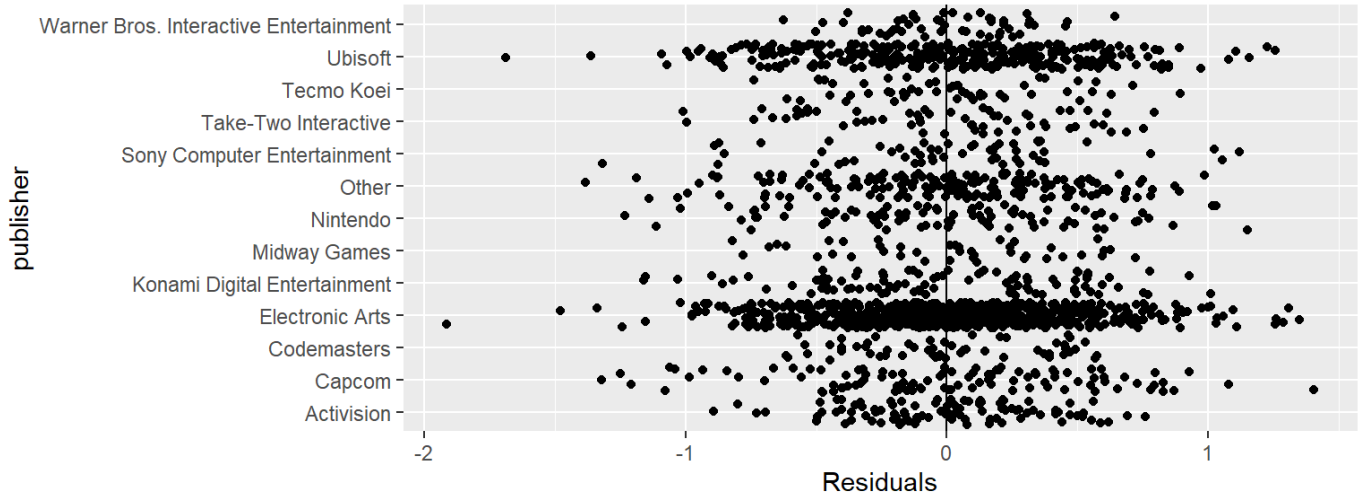
3. Constant variance of Variables

In each of the above charts the categorical variables are showing nice constant variance centered around zero, this linear model has been satisfied.



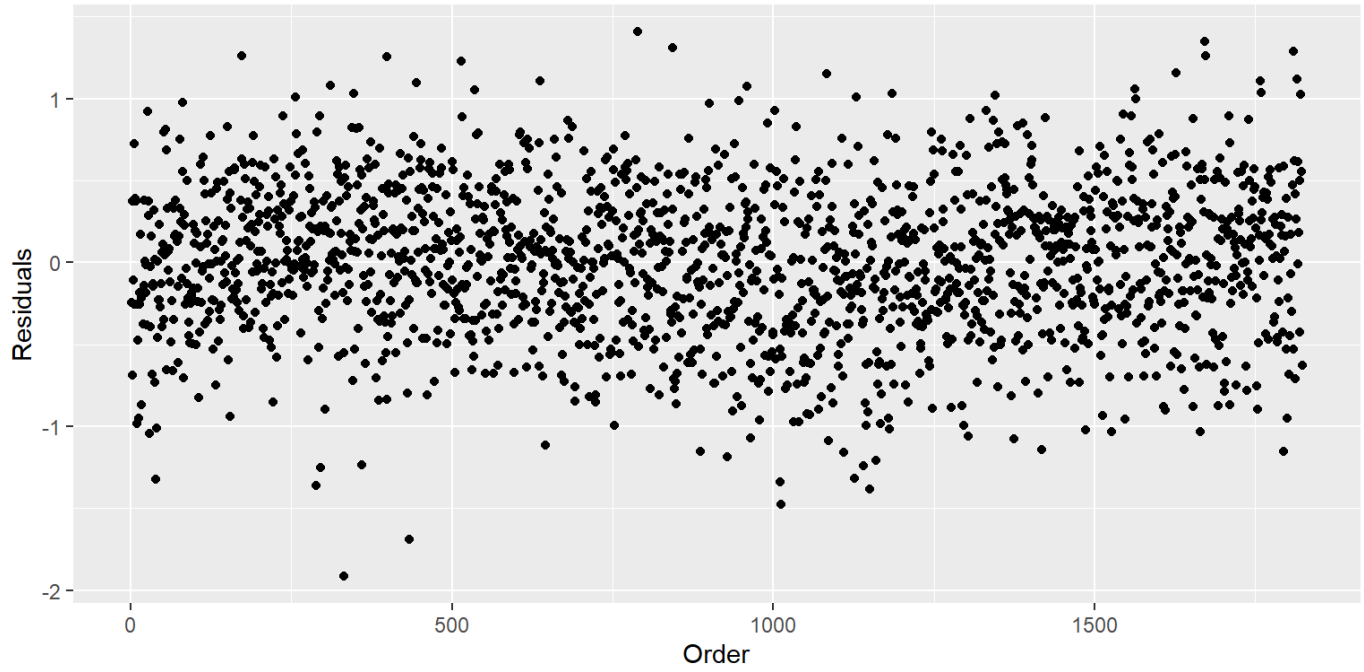






4. Residuals are independent

With samples ordered by year of release, the residuals plotted show no indications of clustering and the data seems evenly distributed around the mean of zero.



Linear Model Results

Using the model tested above we can now proceed to pick the variables which maximise sales. At this point industry knowledge is also critical, specifically we can't choose a games platforms which is obsolete or a publisher that is no longer in business

The model can be used to maximise sales by selecting the options with the highest slopes from each category above. However at this point further analysis needs to be done to check the statistical significance of each of the high scoring category choices. An option may have a high positive slope, but if there isn't enough data to support it, then it should not be part of any recommendation.

Some categories worth considering are listed here:

Genre

- Simulation: 0.1648003150, p-value: 5.511674e-03
- Fighting: 0.1349610929, p-value: 1.412523e-02

Platform

- PS4: 0.3172497448, p-value: 9.333422e-04
- XOne: 0.0577307352, p-value: 5.612009e-01
- WiiU: -0.1194478217, p-value: 2.564689e-01
- 3DS: 0 (baseline)

Publisher

- Nintendo: 0.7429307748, p-value: 8.784536e-23
- Electronic Arts:-0.1252041823, p-value: 2.405702e-02
- Warner Bros: 0.0008069113, p-value: 9.928668e-01

Rating

- M: 0.2885362581, p-value: 2.039512e-08
- E: 0 (baseline)

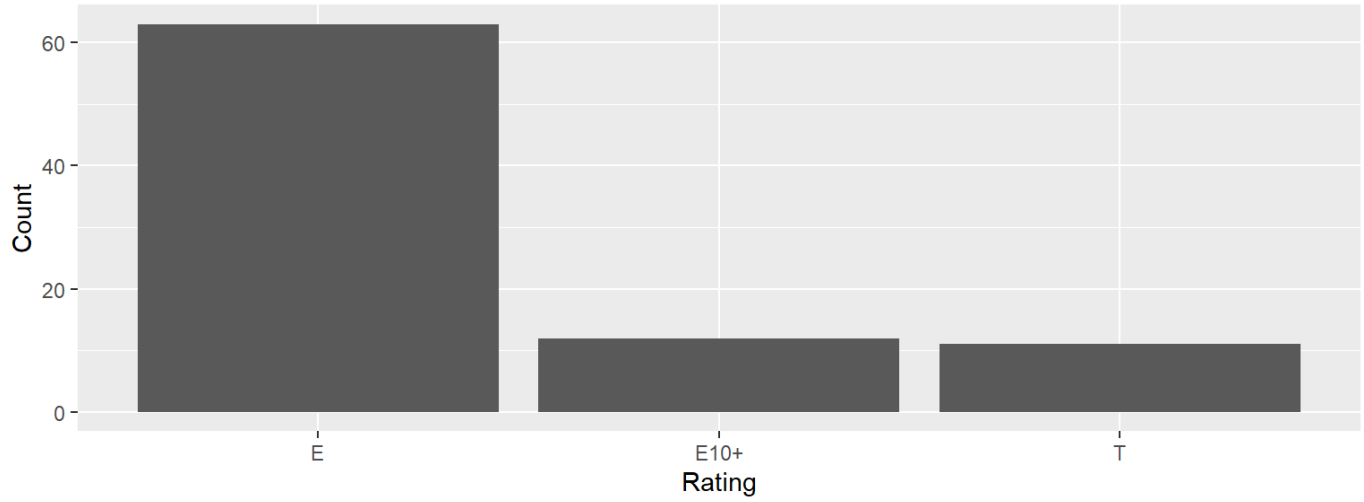
The XOne and WiiU platforms and the Warner Bros publisher all have high p-values so there isn't enough evidence to back up the sales contribution predicted by the model.

The above scores and p-values suggest the following combination:

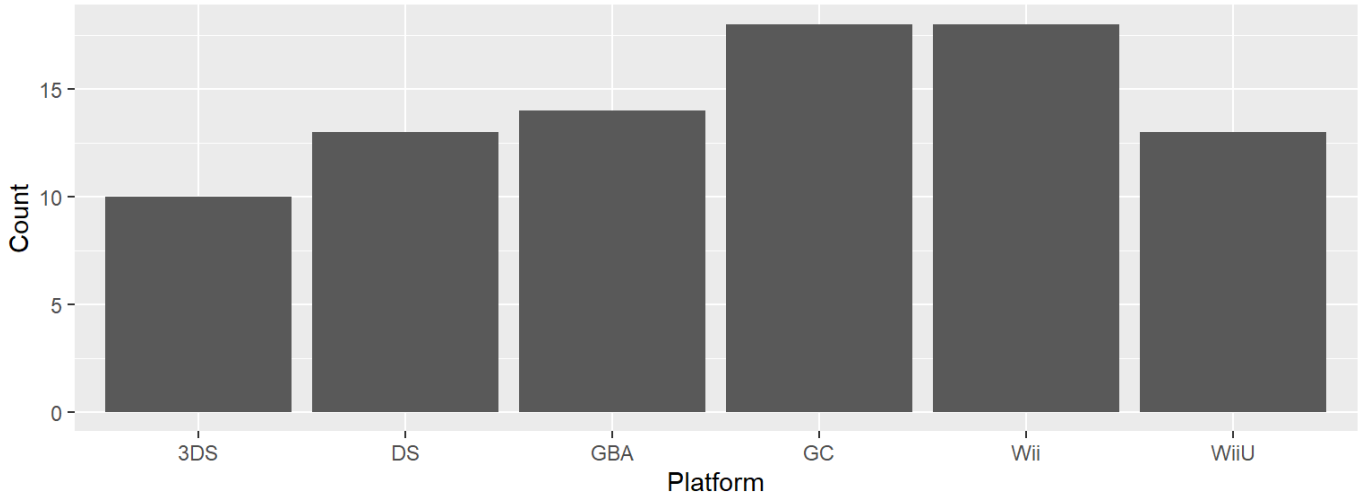
```
10 ^ (
  -0.26742 # Intercept
  + 0.1349610929 # Genre: Fighting
  + 0.3172497448  # Platform: PS4
  - 0.1252041823 # Publisher: Electronic Arts
  + 0.2885362581 # Rating M
) * 1000000
```

> For this combination of choices the model predicts sales of: 2.23 million

There is an alternative approach that could yield even higher sales. The model shows that getting Nintendo as the publisher on average tends to add many millions of to the game sales, so it could be worth trying to develop a game specifically with this in mind. However there are some specific things to bear in mind if targeting Nintendo as a publisher. The first is that they haven't published any games with a 'Mature' rating:



The vast majority of the titles published by Nintendo have a rating of E, suggesting that we should create a game targeting a family friendly audience in order for Nintendo to be more likely to publish. Another constraint here is that Nintendo only publish games for their own platforms:



With this in mind we would need to aim for creating an E-rated game for the 3Ds (which from the model is preferable to the WiiU in 2016). Another possibility is to try and develop for the upcoming Nintendo Switch, but we have no data for this and can't predict how sales will relate to that platform.

```
10 ^ (
  -0.26742 # Intercept
  + 0.1648 # Genre: simulation
  + 0  # Platform 3DS (baseline)
  + 0.7429 # Publisher: Nintendo
  + 0 # Rating E (baseline)
) * 1000000
```

> For this combination of choices the model predicts sales of: 4.36 million

---

T-Tests on interesting linear model outcomes

Games published by Nintendo sell better

A t-test comparing the average sales of games published by Nintendo and games not published by Nintendo yielded the following results:

P-Value	95% CI Lower Bound	95% CI Upper Bound
2.2e-16	-6.198	-5.007

The null hypothesis in the above test is that there really is no difference between the real average sales of games published by Nintendo and the real average sales of games not published by Nintendo. Given how small the p-value is here, the null hypothesis can safely be rejected.

Another interesting point that can be extracted from this t-test is from the 95% confidence interval. According to this test we can say that if this sample were repeated 100 times, then 95 times out of 100, the average sales of games published by Nintendo would be between 5 and 6.2 Million copies more than the average game sales of games not published by Nintendo. This is equivalent to saying:

There is good evidence to say that on average games published by Nintendo sell between 5 and 6.2 million copies more than games not published by Nintendo.

PS4 games sell better than XOne games

A T-test on the average game sales on the PS4 Platform vs the XOne platform yields:

P-Value	95% CI Lower Bound	95% CI Upper Bound
0.0129	0.1333	1.103

From the 95% confidence interval we can say that there is good evidence to say that on average games published on the PS4 will sell between 130,000 and 1.1 million more copies than games on the XOne

Puzzle games average sales vs non puzzle games

A T-test comparing the average sales of puzzle games with non-puzzle games yields the following:

P-Value	95% CI Lower Bound	95% CI Upper Bound
0.4058	-2.099	0.849

The above t-test comparing the sales of puzzle games vs the sales of non-puzzle games has found a high p-value. In this case we don’t enough evidence to support the idea that puzzle games sell more or less copies on average than non-puzzle games.